Evidence of Additional Erroneous Enumerations from the Person Duplication Study

Robert E. Fay, U.S. Census Bureau

Preliminary Version, Oct. 26, 2001

This preliminary version is being made available before the review process used for other ESCAP II reports has been completed. The author will revise the paper on the basis of comments from Census Bureau colleagues. The Census Bureau has decided to release this preliminary version of an analysis the ESCAP considered in making its recommendation of Oct. 16, 2001 not to adjust the census for uses other than redistricting, because of the interest and importance of the findings. In part, the decision to release a preliminary version of this report was to help document the "Revised Early Approximation" of estimated percent undercount presented by Acting Director William Barron and Principal Associate Director for Programs, John H. Thompson, on October 17, 2001. The estimates are preliminary, but the author believes that they will not substantially change with subsequent research.

This preliminary version of the report does not conform to the format of other ESCAP II reports.

Evidence of Additional Erroneous Enumerations
from the Person Duplication Study

Robert E. Fay

ABSTRACT

In January 2001, the U.S. Census Bureau's survey evaluation of Census 2000, the Accuracy and Coverage Evaluation (A.C.E.), estimated a net census undercount of 3.3 million persons. The A.C.E estimate of net census error incorporates estimates of omissions from the census and erroneous census enumerations. If the A.C.E. underestimates erroneous enumerations, it would tend to overstate the net undercount. One evaluation, the Measurement Error Reinterview (MER), estimates that the A.C.E. understates erroneous enumerations by 1.5 million, and a second, the person duplication study, implies that the A.C.E. misclassifies as correct approximately 2.1 million erroneous enumerations due to duplication. To assess their overlap, the study files were merged. The merged results imply an A.C.E. underestimation of 2.6 million erroneous enumerations without allowance for duplicates missed by computer matching. When a uniform allowance for a 75.7% efficiency of computer matching at detecting duplicates is assumed, the additional undetected duplicates imply an A.C.E underestimation of 2.9 million erroneous enumerations. An approximate calculation shows an implied change of approximately 3.2 million in the A.C.E. dual system estimate. The estimates using MER data all make the same implicit assumptions about missing data, and investigation of these assumptions is a topic for future research. The discussion section notes several issues requiring investigation before A.C.E. estimates can be revised on the basis of these preliminary findings.

1. INTRODUCTION

By September 2001, three lines of investigation indicated or suggested that the Accuracy and Coverage Evaluation (A.C.E.) underestimated the number of erroneous enumerations in Census 2000. First, estimates from the Measurement Error Reinterview (MER) (Adams and Krejsa 2001) showed substantial underestimation of erroneous enumerations in the A.C.E. Second, a person duplication study based on computer matching of the A.C.E. E sample to census enumerations (Mule 2001) uncovered a large number of duplicates both to group quarters and to the A.C.E. universe beyond the A.C.E. search area. Comparison of these matched duplicates to their enumeration status in the A.C.E. (Feldpausch 2001a) indicated that the A.C.E. underestimated erroneous enumerations from these sources. Third, A.C.E. estimates of erroneous enumerations of persons living elsewhere on Census Day (April 1, 2000) dropped relative to the 1990 Post Enumeration Survey (Feldpausch 2001b). While the historical comparison is the weakest of the three lines, the apparent drop is consistent with the other two lines of evidence that the A.C.E. underestimated erroneous enumerations.

Previous analyses of the MER and duplicate studies did not address the issue of their overlap. If the estimated erroneous enumerations based on both are simply added, their sum exceeds the estimated net national undercount, 3.3 million persons, from the A.C.E. Logically, the duplication study only measures part of the erroneous enumerations in scope for the MER, so the large result implied by the duplication study suggested that the MER estimate may itself be an underestimate.

To assess their overlap, each study was first individually reviewed and in some cases re-analyzed. Files from the MER and duplication study were merged to permit an assessment of the overlap of the two studies. This approach permits an estimate of the underestimation of erroneous enumerations by the A.C.E. from the combined findings of the MER and duplicate studies.

The MER was based on a reinterview sample drawn within about 1/5 of the sample clusters for the A.C.E. Interviews were conducted primarily in January and February 2001, during the Evaluation Followup (EFU). The original analysis of the MER, termed EFU 1, indicated an underestimation of erroneous enumerations by 1,919,029. To investigate whether the initial study was possibly flawed, a sample was selected from the MER for review by technical specialists under different and more specific coding procedures. No additional data was collected, but both the original A.C.E. data and the EFU interviews were recoded separately and then integrated into a Best Code. Adams and Krejsa (2001, pp. 3-4) revised the estimate of underestimation of erroneous enumerations from 1,919,029 to 1,454,915. The estimate 1,454,915 is itself the net of two estimates: 1,816,315 A.C.E. correct enumerations reclassified as erroneous and 361,400 A.C.E. erroneous reclassified as correct.

The unresolved rate for the EFU 1 analysis was 1.7%. The review increased the unresolved rate for the Best Code to 4.8% and added a new category of Conflicting at 1.0%. The two estimates of 1,919,029 and 1,454,915 make similar implicit assumptions about the treatment of unresolved and conflicting cases, namely to retain the A.C.E. status. Adams and Krejsa (2001, p. 15) provide a table showing a wide range of effects on the overall erroneous enumeration rate resulting from alternative assumptions about the erroneous enumeration rates for unresolved and conflicting cases.

Mule (2001) reported on the results of a computer match of the A.C.E. sample to the census universe, including group quarters. Census 2000 processing for the first time captured names and dates of birth from the forms. There were two stages of matching. The first was based on an exact match of name and date of birth. Census households were provisionally linked together whenever an exact match of one or more persons occurred between them. A second stage was based on statistical matching of any remaining persons in the provisionally linked households. Only some of the second-stage matches were retained as probable matches, on the basis of the similarity of the linked households. Additionally, a subset of exact matches were each evaluated under a model for coincidental agreement of birthday by persons with the same name to estimate a weight reflecting the probability that the match was the same person.

Comparison to A.C.E results indicates that computer matching underestimates total duplication in the census. The A.C.E. estimated duplicates within cluster through clerical matching. Within cluster, computer matching yielded an estimate of only 724,687, 37.8% of the A.C.E. result (Mule 2001, p. 9). In a case-by-case comparison, computer matching discovered few duplicates missed by the clerks *(1)*. An evaluation of the A.C.E. clerical work confirmed its high level of accuracy (Bean 2001, p. 22).

The A.C.E. also used clerical matching to estimate duplication to surrounding blocks in the search area for some cases. Computer matching found an estimated 146,880 duplicates to surrounding blocks, somewhat higher than the A.C.E. result, which was based on searching surrounding blocks for only a subset of A.C.E. cases.

A.C.E. did not separately estimate duplicates beyond the search area. Instead, each duplicated person was assumed to regard one (at most) of these places as the correct residence on Census Day, and to regard the other enumeration as not the Census Day address. The A.C.E. questionnaire specifically attempted to identify persons counted at an incorrect address in order to include them in the estimated erroneous enumerations. Mule (2001, p. 11) reported the estimate of 2,089,107 duplicated persons outside the surrounding blocks to the housing unit population and additionally 660,094 to group quarters *(2)*. The evidence from the success of computer matching within cluster compared to clerical suggests that both figures are underestimates. In fact, if the 37.8% efficiency observed by Mule were to apply to these estimates as well, then the total number of duplicates to group quarters and to housing units beyond the search area would be considerably larger than implied by the A.C.E. estimates of persons enumerated at an incorrect address.

An analysis of Mule's results presented here suggests an alternative measure of efficiency. When the success of computer matching is measured not only against the A.C.E. population but also the housing units reinstated or deleted during the Housing Unit Duplicate Operations in Census 2000, which also used computer matching, then computer matching found a proportion of duplicates higher than 37.8%. An analysis presented here suggests that the efficiency of computer matching within cluster was possibly as high as 75.7%. But this estimate is probably an upper bound on the efficiency for identifying duplicates beyond the search area or to group quarters.

Feldpausch (2001) compared the duplicates identified by computer matching outside the search area to their classification in A.C.E. She noted that the erroneous enumeration rates were considerably lower than expected for these cases, but her report did not summarize the findings in terms of their implications for the A.C.E. estimate of erroneous enumerations. Approximate calculations presented here indicate that the A.C.E. missed an estimated 498,070 erroneous enumerations from duplication with group quarters and 1,555,090 to the household population, for a total of 2,053,160. The estimate of 2,053,160 A.C.E. correct reclassified as erroneous for reasons of duplication logically should have been only part of the MER estimate of 1,816,315 A.C.E. correct reclassified as erroneous. In the extreme case that MER had caught none of these erroneous enumerations due to duplicates, adding the estimate of 2,053,160 from the duplication study, plus 1,454,915 from the MER, would exceed the estimated undercount from the A.C.E. The sum fails to account for the overlap between the studies, however.

To assess the overlap, computer files from the duplication study and the MER review sample were merged. The result confirmed that the MER found some of the duplicates missed by the A.C.E., but it failed to detect others. Using the estimation procedure implicit in the current estimate of 1,454,915 from the MER, an additional 1,125,377 should be added, to bring the total from the two studies to 2,580,292. But this estimate does not reflect the efficiency of computer matching at finding duplicates. If the estimate of 75.7% based on the within cluster efficiency in the duplication study is used, then the estimate of the additional component, 1,486,257, raises the estimate of erroneous enumerations missed by the A.C.E. to 2,941,172.

2. THE MEASUREMENT ERROR REINTERVIEW (MER)

The E sample component of the A.C.E., a sample of census enumerations, was used to estimate erroneous enumerations. Erroneous enumerations can be broadly categorized into *duplicates, fictitious enumerations, geocoding errors, other residence,* and *persons with insufficient information for matching*. *Duplicates*, that is, duplicates within the sample clusters or surrounding blocks, were identified clerically, as were persons with insufficient information to match. Determination of *geocoding errors* required interviewers to find the census units but not necessarily to contact the residents. Identification of *fictitious enumerations* and *persons with other residence* on Census Day required collection of information from respondents who were frequently, but not always, members of the original census household. The other component of the A.C.E., an independently listed and interviewed P sample, was used to estimate census omissions.

In both the 1990 Post Enumeration Survey and the 2000 A.C.E., there was a residual category of *unresolved* cases. In both studies, a probability of correct enumeration was imputed for these cases without categorizing them by type of erroneous enumeration. In 1990, the overall rate of erroneous enumeration was 5.74%, including an imputed contribution of 0.26% from *unresolved* (Feldpausch 2001b, p. 20). In 2000, the A.C.E. estimated a rate of 4.72%, including a contribution of 0.62% for *unresolved*. Much of the larger contribution in 2000 from *unresolved* can be attributed to stricter rules in 2000 for *other residence* Feldpausch (2001b, pp. 18-19). The contribution of *other residence* was 2.18% and 1.03%, in 1990 and 2000, respectively.

Feldpausch (2001b, pp. 20-21) redistributed *unresolved* in 1990 and 2000 into their likely categories, increasing the contribution of *other residence* to 2.29% and 1.41% in 1990 and 2000, respectively (Table 1). The redistribution lessens the gap between 1990 and 2000, but the difference between 1990 and 2000 in the estimates is about .88% (that is, .88 percentage points). In other words, the 2000 A.C.E. appears to have estimated approximately 2.3 million fewer erroneous enumerations from *other residence* than it would have at the 1990 rate.

Table 1 Comparison of percent erroneous enumerations by type from the 1990 PES and the 2000 A.C.E. after re-classifying the unresolved people. For comparability, the PES rates were computed for the A.C.E. universe.

	1990 PES	2000 A.C.E.
Duplicate	1.66%	0.76%
Fictitious	0.22	0.50
Geocoding Error	0.38	0.25
Other Residence	2.29	1.41
Insufficient Information	1.19	1.80
Total	5.74%	4.72%

Source: Feldpausch (2001b, p. 21).

Besides the decline in *other residence,* there are other components with large changes. The changes in these components have been evaluated by other studies *(3).*

The A.C.E. E-sample interviewing occurred in two waves. A first wave corresponded to the initial interviewing for the P sample, accomplished by telephone (April 24 through June 13, 2000, 29.4% of total workload) and personal visit (June 18 through Sep. 1, except for one FL office completed by Sep. 11) (Hogan 2001, p. 25; Bryne et al. 2001, pp. 3, 7), using the same computer-assisted instrument. The initial interview determined both the current and Census Day residents. When P-sample Census Day persons were matched to the census and classified as residents, the corresponding census enumerations were classified as matched in the E sample and classified as correct census enumerations. The remaining census enumerations selected in the E sample were sent in Oct. and Nov. 2000 to A.C.E. Person Followup (PFU). The PFU used a paper questionnaire. For the nonmatched E-sample cases, the PFU attempted to determine if the census enumerations represented valid Census Day residents.

The Measurement Error Reinterview (MER) was designed to evaluate both the P- and E-sample data collection. A sample of approximately 1/5 of the A.C.E. clusters was selected, followed by subsampling of housing units within the clusters sampled for the MER. With respect to the E sample, the MER's principle objectives were to detect errors in the A.C.E. measurement of *fictitious enumerations*, *geocoding errors*, and *other residence*. The MER sample was reinterviewed with a paper questionnaire during Evaluation Followup (EFU) in Jan. and Feb. 2001.

Initial clerical coding of the EFU results, later termed *EFU 1*, was partially supplemented with review by senior analysts. Interviewer notes were frequently consulted to classify cases. If the EFU results were incomplete or questionable, coders could reject the EFU results and substitute the A.C.E. results. On a weighted basis, approximately 6.4% of the EFU results were rejected

(Adams and Krejsa, 2001, p. 12). The initial analysis of the EFU 1 estimated an A.C.E. understatement of erroneous enumerations of 1,919,029.

In response to a request from the Evaluation Steering Committee for A.C.E. Policy (ESCAP), Martin (2001a, 2001b) reviewed the paper questionnaires used for PFU and EFU. Her report detailed limitations with both questionnaires. The EFU questionnaire failed to asked the Census Day address for a number of situations in which the information would be required to make an unambiguous determination of erroneous enumeration *(4).* This logical flaw could have reflected the lack of field-testing of the EFU questionnaire prior to its use, unlike the PFU questionnaire, which was field-tested (Martin 2001a, p. 2). In several other respects, however, the EFU questionnaire was superior in design to the PFU for clarifying concepts, and for reminding respondents by probing for details. Martin (2001a, pp. 9-10) judged both questionnaires as substantial improvements over the PES and evaluation questionnaires in 1990. On the other hand, she remarked that the extensive reliance on interviewer notes to code both the PFU and EFU questionnaires meant that the questions failed to reflect many persons' circumstances in a structured way *(5).* Both questionnaires were challenged by the complexity of census residency rules (Martin 2001a, pp. 10-11).

A subsample of the EFU, the Person Followup/Evaluation Followup Review (Adams and Krejsa 2001), was selected for systematic recoding by analysts, who were the staff most highly experienced in the coding phase of coverage studies. New coding procedures were established. The original A.C.E codes, now identified as *PFU 1,* were reviewed and revised as appropriate using only the A.C.E. responses, to yield *PFU 2.* Similarly the EFU responses were recoded as *EFU 2,* using only EFU responses. Coders then combined the information into a *Best Code*. One of the guiding principles was to favor information from nonproxy respondents, who were members of the original E sample household, over information from proxy respondents, who were not. Some types of proxy respondents were regarded as more informed than others. Other rules also attempted to follow the principle of using the information that appeared the most complete (Adams and Krejsa, 2001, p. 3). A *Conflicting* category was allowed for cases with contradictory geographic information or contradictory information on Census Day residence and the same type of respondent; that is, either both non-proxies or both proxies considered equally informed. The EFU 2 coding recognized the logical flaw in the EFU questionnaire, coding cases with another Census Day address that was not recorded to unresolved instead of erroneous as in EFU 1 (Adams and Krejsa 2001, pp. 2, 3, 13). To analyze the results, Adams and Krejsa compared the PFU 1 code to the Best Code (Table 2).

Table 2 Comparison of Production (PFU 1) to Best Code from the PFU/EFU Review. For PFU 1, total correct enumerations are divided into those matched to P-sample residents and nonmatched correct enumerations in the census. Estimates are weighted by MER review weights, reflecting A.C.E. sampling, MER subsampling, and subsampling for the review sample.

	Best Code from Second Review				
PFU 1 Code	Correct	Erroneous	Unresolved	Conflicting	Total
Total Correct Enumerations	238,786,314	1,816,315	9,151,011	1,613,442	251,367,081
Matched	210,222,189	1,139,407	8,763,973	563,514	220,689,083
Nonmatched Correct Enum.	28,564,125	676,908	387,038	1,049,928	30,677,998
Erroneous	361,400	2,936,887	186,418	666,512	4,151,217
Unresolved	2,529,422	664,929	3,303,074	314,685	6,812,110
Total	241,677,134	5,418,131	12,640,503	2,594,639	262,330,408

Source: Adams and Krejsa (2001, p. 7).

Based on Table 2, Adams and Krejsa reported an estimate of the understatement of erroneous enumerations by the A.C.E. of 1,454,915. To represent their estimator, let $M_{i,j}$ denote the estimated frequencies in Table 2, where i denotes PFU 1 outcomes C, E, U *or* "**.**" for total correct enumerations (both matched and nonmatched correct enumerations), erroneous, unresolved, and total, respectively, and j denotes Best Codes C, E, U, Cf or "**.**" for correct, erroneous, unresolved, conflicting, and total. The estimate of 1,454,915 is the difference of two cells in Table 2: $M_{C,E}$ = 1,816,315 A.C.E. correct enumerations reclassified as erroneous, and $M_{E,C}$ = 361,400 A.C.E. erroneous reclassified as correct. Their estimator is

$$B_{MER} \quad = \quad M_{C,E} \ - \ M_{E,C} \tag{1}$$

Estimating the net error in A.C.E. according to eq. (1) implicitly uses the remaining information in Table 2 in the following manner:

C $M_{C,U}$ = 9,151,011 and $M_{C,Cf}$ = 1,613,442 correct enumerations in PFU 1 classified as unresolved or conflicting, respectively, by the Best Code are in effect treated as correct, since this information is not reflected in eq. (1);

C $M_{E,U}$ = 186,418 and $M_{E,Cf}$ = 666,512 erroneous enumerations in PFU 1 classified as unresolved or conflicting, respectively, by the Best Code are in effect treated as erroneous, since this information is also not reflected in eq. (1); and

C A.C.E. imputations for $M_{U,.}$ = 6,812,110 are accepted without modification by the information from the Best Code.

In short, the estimated net error is based on the only two cells, $M_{C,E}$ and $M_{E,C}$, in which a Best Code of erroneous or correct revises an A.C.E. one with the opposite value.

Of the $M_{C,E}$ = 1,816,315 A.C.E. correct enumerations reclassified as erroneous, an estimated 614,451, approximately 1/3, were because the person should have been counted in a group quarters, including 315,406 in dorms (Adams and Krejsa 2001, p. 9). An additional 976,343 were attributed to other residence issues, 28,729 to *fictitious*, 120,530 to *geocoding*, and 76,262 for other reasons *(6)*. The estimate for other residence issues includes 73,940 for *joint custody*.

Adams and Krejsa (2001, p. 15) remark on the sensitivity of their findings to alternative assumptions about the unresolved and conflicting cases. Overall, an estimated 9,337,429 cases coded as correct or erroneous in PFU 1 became unresolved ($M_{C,U}$ + $M_{E,U}$). Unlike EFU 1, unresolved cases in EFU 2 include cases with another Census Day address that was not recorded because of the logical error in the EFU form, and these in turn contribute to unresolved for Best Code. These cases are largely analogous to the group of cases in the A.C.E. E sample requiring imputation of the probability of correct enumeration: those with another Census Day residence, where the address had not been recorded (Feldpausch 2001b, pp. 18-19, cited previously; also Cantwell and Childers 2001). In A.C.E., the unresolved in this category were imputed an estimated probability of correct enumeration of approximately 0.23, much lower than most other categories (Cantwell et al. 2001, p. 42). Applying a similar approach to unresolved EFU 2/Best Code outcomes could increase the estimated erroneous enumerations, compared to the assumptions implicit in eq. (1).

An estimated 2,279,954 cases coded as correct or erroneous in PFU 1 became conflicting ($M_{C,Cf}$ + $M_{E,Cf}$). Conspicuously, these are disproportionately drawn from PFU 1 nonmatched correct enumerations and erroneous enumerations.

3. THE PERSON DUPLICATION STUDY

As noted previously, the A.C.E. estimate of duplicates within block is quite accurate (Bean 2001, p. 22). The A.C.E. estimate of duplicates in Table 1 also includes a smaller number of duplicates to surrounding blocks in the area of search.

Beyond area of search, however, the A.C.E. measured duplicate enumerations indirectly, as part of two other categories of erroneous enumerations. The same strategy was used by the 1990 PES. The first category is *geocoding error*: If the census incorrectly assigns an address to a block outside of the area of search, then the case is classified as a *geocoding error*. The census sometimes duplicates a unit by including it in both an incorrect block and the correct one. In the A.C.E. estimation, a duplicated unit is counted as an erroneous enumeration due to *geocoding error* when the enumeration in the wrong block is selected in the E sample, and as a correct enumeration when the correct block is selected. The second category is *other residence*: If a duplicate occurs between a correct enumeration and an incorrect one at the wrong residence, the person is counted as correct when the correct address is selected, and as erroneous when the wrong residence is selected. As previously recognized *(7),* estimates of *other residence* include both duplicate enumerations of persons enumerated correctly at their Census Day address, and persons who are only counted at the wrong address.

A person duplication study was initiated in 2001 in part to investigate issues arising from the Housing Unit Duplication Operations (Nash 2000, Fay 2001), which were part of census operations in 2000. The Housing Unit Duplication Operations attempted to correct problems in the development of the Master Address File (MAF) for the census—the MAF duplicated some housing units, often with similar but not identical address descriptions. The duplication of housing caused the duplicate enumeration of persons in early census returns. As an example, when the MAF had included separately both “Apt 1A” and “Unit A1” at the same street address, the early census returns had often counted the same households.

The Housing Unit Duplication Operations were divided into two phases for purposes of timing, particularly with respect to the A.C.E. In the first phase, addresses were linked either by rules applied to the address entries in the MAF or by computer matches of housing units apparently containing the same people. Person matching was carried out in two steps: a first step of finding exact matches of persons with the identical first and last names, and month, day, and years of birth, and a second step of evaluating the similarities of households linked by one or more exactly matching people. Person matching was restricted to households in the same state and within 30 miles of each other. For each pair of units linked either by address edits or person matching, one unit was selected to remain in the census, and the other was removed from the A.C.E. universe *(8)*.

In the second phase, more detailed criteria were developed and applied to determine which of the units removed in the first phase would be reinstated into the census. Phase 2 reinstated 2,315,553 persons and 1,002,951 housing units into the census, and deleted the remaining 3,572,799 persons and 1,371,320 housing units. The reinstated units were included in the census count but excluded from the A.C.E. universe. Although the number of reinstated persons is large, Raglin (2001) showed that the effect on the A.C.E. dual system estimation of total population was minor.

In spring 2001, a team of Census Bureau staff began a study of computer matching of the A.C.E. sample to the entire census nationally (Mule 2001). The study had multiple objectives, including to account for the drop in estimated *duplicates* from 1990 to 2000 (Table 1) and to investigate the quality of reporting of *other residence* in the A.C.E. A source file comprising the A.C.E. E-sample universe and reinstated units in the A.C.E. sample clusters was matched to a target file comprising the enumerated persons in census housing units (except for reinstated units), group quarters, reinstated units, and deleted units (even though deleted units are excluded from the final census).

Computer matching methods were similar to those used during the Housing Unit Duplication Operations, but they incorporated several refinements. The first stage was again based on exact matching, but this time the matching also considered middle initials. Some errors in the capture of names were repaired prior to the match *(9)*, and some combinations of names and birthdays identified and removed from analysis, including the tendency, particularly among the Hispanic population, to name a child born on an important Saint's Day after the saint *(10)*.

The second stage of matching considered households linked together by one or more exact matches. Possible matches between the remaining persons were evaluated with Census Bureau matching software based on the Fellegi-Sunter algorithm. Some of the exact matches from the first stage were evaluated for coincidental sharing of birthday. Because some combinations of first and last names are common, such as “Linda Smith,” a model weight was developed to adjust for the probability of two persons with the same name sharing a birthday (Mule 2001, pp. J-3 to J-4). Model weights of 1 or 0 were assigned for the remaining cases according to Table 3.

The team developed rules to reduce the chance of false match by dropping a number of preliminary matches from the analysis. Greater evidence was required for matches between different states than within (Table 3). The rules permitted single individuals to match, including to group quarters, through first-stage matches only.

Table 3 Rules for retaining matches based on the number of persons linked between households, stage, and geography. The category with weight 0 was dropped from the analysis.

Category	Stage	Model Weight
Same State		
2+ links, 1-1 match of households	1st or 2nd	1
2+ links, not all persons matched	1st or 2nd	1
1 link only (including to group quarters)	1st	model
Different State		
2+ links, 1-1 match of households	1st or 2nd	1
2+ links, not all persons matched	1st	model
2+ links, not all persons matched	2nd	0
1 link only (including to group quarters)	1st	model

Source: Table J2, p. J-2, Mule (2001). A category involving only 6 cases with links within the same housing unit is omitted here.

The estimates were weighted by the product of A.C.E. sampling weights, the model weights for coincidental birth, and multiplicity weights. The multiplicity weights adjusted for multiple chances of selection, where appropriate. For example, when an A.C.E. household duplicated another household not in the selected A.C.E. sample, a multiplicity weight of ½ was applied since either household in the pair was eligible for selection into the A.C.E. (Mule 2001, p. 6 and App. G). Multiplicity weights were not required in other instances, such as duplication to group quarters, since group quarters were not sampled for the A.C.E. Table 4 summarizes the resulting estimates.

Table 4 Estimated number of computer matches from the 2001 computer match study of census person duplicates. Reinstated and deleted units were determined by the Housing Unit Duplication Operations; neither are included in the A.C.E. universe, and deleted units are not included in the final census. Estimates are weighted by A.C.E. sampling weights, multiplicity weights, and weights computed from the Poisson model for coincidental birth.

	Within Cluster	Surrounding Blocks	Beyond Search Area	Total
E-sample Elig. to E-sample Elig.	724,687	146,880	2,089,107	2,960,675
E-sample Elig. to Reinstated	1,049,699	24,029	573,698[a]	1,647,426[a]
E-sample Elig. to Deleted	1,941,732	682,909	276,968[b]	2,901,609[b]
E-sample Elig. to Group Quarters	103,168	46,736	510,190[c]	660,094[c]

Source: Mule (2001) Table 2 (p. 9) and Table 6 (p. 11). Cells are noted where the estimate shown may contain small numbers of matches of reinstated to reinstated, deleted, and group quarters beyond the search area. The table was derived from the tables in his report.
Notes: a. Includes a presumably small number of duplicates of reinstated to reinstated beyond the search area.
b. Includes a presumably small number of duplicates of reinstated to deleted beyond the search area.
c. Includes a presumably small number of duplicates of reinstated to group quarters beyond the search area.

The large number of duplicates identified between the A.C.E. E sample and reinstated and deleted units reflects the basic design of the Housing Unit Duplication Operations, which used computer matching to identify possible duplicates. Mule (2001, pp. 9-10) showed that if duplicates to reinstates and group quarters within the search area were added to the A.C.E. estimate of duplicates to the E-sample eligible universe, then the outcome would have been considerably closer to 1990 levels of duplication, and that if duplicates to deleted units were also added, the result would have exceeded 1990 levels.

Because duplicates beyond the search area are included in other categories of erroneous enumeration, it is only possible to directly compare A.C.E. clerical performance with computer matching within the search area (Table 5). As previously noted, Mule (2001, p. 9) reported that computer matching within cluster found only 37.8% the number of duplicates as the A.C.E. But if all enumerated census housing units are considered—that is, before the Housing Unit Duplication Operations had deleted some units and removed the reinstated units from the A.C.E. universe—computer matching would have matched an estimated 3.7 million persons within cluster. Assuming also that clerical matching would have found no additional matches to the reinstated or deleted units, then computer matching may have been as much as 75.7% *(11)* efficient at finding duplicates within cluster. To the extent that clerical matching may have found some additional matches to the reinstated or deleted units, the efficiency would be somewhat lower.

Table 5 Comparison of A.C.E. and Computer Match Study of Census Duplicates Within the A.C.E. Search Area. Estimates are weighted A.C.E. sampling weights, multiplicity weights, and weights computed from the Poisson model for coincidental birth.

	Within Cluster	Surrounding Blocks	Total Within Search Area
A.C.E. clerical estimate	1,916,340[a]	98,335[a]	2,014,675
Computer: E-sample Elig. to E-sample Elig.	724,687	146,880	871,567
Computer: E-sample Elig. to Reinstated	1,049,699	24,029	1,073,728
Computer: E-sample Elig. to Deleted	1,941,732	682,909	2,624,641
Total Computer	3,716,118	853,818	4,569,936
Computer: E-sample Eligible to Group Quarters	103,168	46,736	149,904

Source: Mule (2001) Table 2 (p. 9), except where noted.
Note: a. Computed by Thomas Mule but not shown separately in his report.

In addition to ignoring possible additional clerical matches to the reinstated or deleted households, there are other reasons that 75.7% efficiency would overstate the efficiency of computer matching outside of the search area or to group quarters. The rules in Table 3 favor the matching of whole households, generally the situation addressed by the Housing Unit Duplication Operations, but they are more cautious in retaining partial households as confirmed matches. Matches to different states require more evidence than within, and matches to group quarters or of single persons only require matching at the first stage. Thus, the estimate of 75.7% is interpreted in subsequent calculations as if it were an upper bound on the efficiency of computer matching.

4. COMPARING CENSUS DUPLICATION TO THE CORRESPONDING A.C.E. ENUMERATION STATUS

Feldpausch (2001) analyzed the A.C.E. Enumeration Status from the duplication study. Because Mule (2001) had primarily analyzed the results within the search area, Feldpausch concentrated on matches beyond the search area. She tabulated duplicates of A.C.E. sample cases from the person duplication study to the A.C.E. universe and group quarters. She distinguished between group quarters where persons could claim that they had a usual home elsewhere (UHE), such as military barracks and soup kitchens, from others, such as college dormitories, nursing homes, and jails, where census rules required that the residents be counted at their group quarters. To restrict the analysis to likely duplicates, she included only cases with model weight > 0.5 based on the Poisson model (Table 6).

Table 6 Estimated E-sample duplicates to enumerations outside the search area. Estimates for group quarters are distinguished by whether a usual home elsewhere (UHE) could be reported. Estimates are weighted A.C.E. sampling weights, multiplicity weights, and weights computed from the Poisson model for coincidental birth. Only duplicates with weight from Poisson model > 0.5 are included. Imputed probabilities of correct enumeration are used to apportion cases where the enumeration status was unresolved.

		Group Quarters			
		Cannot Claim UHE		Can Claim UHE	GQ Total
Final Match Code	E-Sample Eligible HU	Dorm	Other		
Total Correct Enumerations	1,862,228	147,901	158,436	52,139	358,477
Matched	1,298,084	93,846	103,871	40,378	238,095
Nonmatched Correct Enum	564,144	54,055	54,565	11,761	120,382
Erroneous	307,138	123,257	31,320	7,447	162,024
Total	2,169,366	271,158	189,756	59,586	520,501

Source: Feldpausch (2001) Table 1, p. 4, and Table 3, p. 6.

An approximate estimate of how much the A.C.E. understates erroneous enumerations, most of which should have been *other residence*, can be based on the results in Table 6. An estimated 2,169,366 duplicates to the housing unit population appears offset by only 307,138 estimated

A.C.E. erroneous enumerations. The estimation of 307,138 uses multiplicity weights, generally 1/2. Thus, 307,138 should be multiplied by approximately 2 to compensate for the effect of multiplicity weighting, in order to approximate the contribution of these cases to the estimated erroneous enumerations in the A.C.E. (In fact, the factor should be slightly greater than 2 for a different reason, to compensate for the model weights reflected in the table. In fact, most of these weights were close to 1.) Consequently, the results imply approximately 2,169,366 - 2(307,138) = 1,555,090 undetected erroneous enumerations from duplication to the household universe.

Duplicates to the group quarters where persons can claim UHE may be interpreted in two ways. Feldpausch (2001, p. 4) interprets this group as correct enumerations in the A.C.E. universe, so that the erroneous enumeration might be assigned instead to the group quarters universe outside of the scope of the A.C.E. Although the group is smaller than others, the analysis here includes these duplicates in the erroneous enumeration total for two reasons. First, although UHE is allowed for these cases, it is not required, and there may be some of these persons who considered themselves residents of the group quarters. Second, the implicit strategy to estimate the population with the A.C.E. is the following:

$$\begin{gathered}\textit{Corrected U.S. Population} = \textit{Corrected population in A.C.E. universe} + \\ \textit{census count outside the A.C.E. universe}\ .\end{gathered} \tag{2}$$

The strategy implicitly relies on previous evidence that the net error of the second term would be small. An unexpectedly high duplication rate between the A.C.E. universe and the balance represents an issue for eq. (2), regardless where the error from duplication is assigned. Consequently, the analysis in this section and the next includes duplicates to group quarters where UHE is permitted, even though this approach favors eq. (2) over the separate question of how well the A.C.E. measured the A.C.E. universe.

Although Feldpausch limits her analysis to duplicates outside of the search area, all duplicates to group quarters are of interest, including those within the search area. The number of duplicates missed by the A.C.E. can be approximately estimated as the difference, 498,070, of the total number of duplicate enumerations to group quarters, 660,094 (Table 4), and the estimated number of erroneous enumerations in Table 6, 162,024. (Unlike duplication to the housing unit population, the issue of multiplicity weighting does not arise because group quarters were excluded from the A.C.E. universe. This approximation overlooks the possibility that some A.C.E. duplicates to group quarters within the search area may have been classified as erroneous for other reasons. The analysis in the next section avoids this assumption.)

5. THE MERGED MER AND PERSON DUPLICATION STUDIES

To interpret the joint implications of the MER review and person duplication studies, their files were merged *(12)*. The merged file was restricted to the MER review sample, and MER review weights were used in the analysis. For simplicity, only duplicates with model weights greater than .98 under the Poisson model for coincidental births were included *(13)*. Tables 7 mirrors Table 2 but displays MER review sample cases identified as duplicates to group quarters.

Table 7 Comparison of Production (PFU 1) to Best Code from the PFU/EFU Review for persons matched to the group quarters universe in the person duplication study, with model weight from the Poisson model > .98. Estimates are weighted with MER review weights.

	Best Code from Second Review				
PFU 1 Code	Correct	Erroneous	Unresolved	Conflicting	Total
Total Correct Enumerations	93,057	54,849	24,507	28,782	201,195
Matched	67,036	27,756	24,507	0	119,299
Nonmatched Correct Enum	26,022	27,093	0	28,782	81,896
Erroneous	0	118,812	491	8,647	127,950
Unresolved	5,901	2,648	7,608	1,965	18,122
Total	98,958	176,310	32,606	39,393	347,267

Restricting Table 7 to cases with model weight > .98 produces a lower estimated total, 347,267, compared to 520,501 in Table 6, although the difference is also due to the different samples in the two tables. Using the notation of Section 2, let $DG_{i,j}$ denote the estimated duplicates to group quarters in Table 7. Approximately 1/3 of the total, $DG_{E,E}$ = 118,812, is consistently identified as erroneous by PFU 1 and the Best Code. The MER estimate of bias, eq. (1), includes $DG_{C,E}$ = 54,849. But the estimator fails to capture the following information as evidence provided by the duplicate study of further A.C.E. underestimation of erroneous enumeration:

C $DG_{C,C}$ = 93,057 PFU 1 Correct to Best Correct
C $DG_{C,U}$ = 24,507 PFU 1 Correct to Best Unresolved
C $DG_{C,Cf}$ = 28,782 PFU 1 Correct to Best Conflicting
C $DG_{E,C}$ = 0 PFU 1 Erroneous to Best Correct

The total is 146,346 additional duplicates to group quarters, unmeasured by PFU 1 and the estimator given by eq. (1). Formally, the estimator of additional bias from duplicates to the group quarters population may be written

$$ABG_{DUP} \quad = \quad DG_{C,C} + DG_{C,U} + DG_{C,Cf} + DG_{E,C} \qquad (3)$$

This estimator, like eq. (1), does not alter the assumptions made about unresolved PFU 1 cases, namely, that the A.C.E. imputation procedures represent the correct proportion of erroneous enumerations among unresolved PFU 1 cases.

Similarly, Table 8 mirrors Table 2 for duplicates to the E-sample eligible population. Three factors account for differences between total duplicates in Table 6 and Table 8: 1) the use of the full A.C.E. sample in Table 6 and the MER review sample in Table 8; 2) more restrictive conditions on the model weight in Table 8; and 3) the use of multiplicity weights in Table 6, generally ½ for matches to E-sample eligible housing units. The combination of the last two reasons accounts for an estimated total in Table 8 somewhat less than twice that in Table 6.

Table 8 Comparison of Production (PFU 1) to Best Code from the PFU/EFU Review for persons matched to the E-sample eligible universe in the person duplication study, with model weight from the Poisson model > .98. Estimates are weighted with MER review weights, but do not include a multiplicity weight.

	Best Code from Second Review				
PFU 1 Code	Correct	Erroneous	Unresolved	Conflicting	Total
Total Correct Enumerations	2,667,580	191,047	162,181	94,244	3,115,053
Matched	2,123,795	131,842	143,875	4,500	2,404,013
Nonmatched Correct Enum.	543,785	59,205	18,306	89,744	711,040
Erroneous	2,548	391,371	9,866	58,405	462,190
Unresolved	69,669	44,889	168,205	35,037	317,801
Total	2,739,797	627,307	340,252	187,687	3,895,044

There are 3,895,044 total duplicates without considering multiplicity. Only about half of these, 1,947,522, should be correct enumerations. The following cells are considered to be correct enumerations in eq. (1):

C $DE_{C,C}$ = 2,667,580 PFU 1 Correct to Best Correct
C $DE_{C,U}$ = 162,181 PFU 1 Correct to Best Unresolved

C $DE_{C,Cf}$ = 94,244 PFU 1 Correct to Best Conflicting
C $DE_{E,C}$ = 2,548 PFU 1 Erroneous to Best Correct

The total of these components is 2,926,553, implying an underestimation of erroneous enumerations of 979,031 from undetected duplicates within the A.C.E. universe, in addition to duplicates identified in MER. Formally, the estimator of additional bias from duplicates to the E-sample eligible population may be written

$$ABE_{DUP} \;=\; DE_{C,C} \;+\; DE_{C,U} \;+\; DE_{C,Cf} \;+\; DE_{E,C} \;-\; .5\, DE_{.,.} \qquad (4)$$

It may also be written

$$ABE_{DUP} \;=\; .5\, DE_{.,.} \;-\; DE_{E,E} \;-\; DE_{E,U} \;-\; DE_{E,Cf} \;-\; DE_{C,E} \;-\; DE_{U,.} \qquad (5)$$

As in eq. (3), the estimator assumes that the A.C.E. imputation procedures represent the necessary proportion of erroneous enumerations among the unresolved.

To represent the possible effect of the efficiency of computer matching, each term of eq. (3) and (4) was assumed proportionately understated. As a logical constraint, the effect of adding the adjusted estimates from Tables 7 and 8 should not exceed the corresponding entries of Table 2, and this constraint is not violated for the assumed 75.7% efficiency.

Table 9 summarizes the separate contributions of eq. (1) from the MER review sample and the sum of eq. (3) and (4). The table shows results for Non-Hispanic Blacks, for Hispanics, and for Non-Hispanic Whites and all other races separately, classified according to the A.C.E. poststratification.

Although revised dual system estimates in full detail have not been computed to reflect the findings of Table 9, the effect of the estimated 2,941,172 extra erroneous enumerations on the A.C.E. dual system estimate will exceed 3 million. Davis (2001, p. 5) summarizes the A.C.E. dual system estimate, *DSE,* as

$$DSE \;\;'\;\; DD \times \frac{CE}{N_e} \times \frac{N_p}{M} \qquad (6)$$

where

DD = the number of census data-defined persons eligible and available for A.C.E. matching;
CE = the estimated number of correct enumerations from the E Sample;
N_e = the estimated number of people from the E Sample;
N_p = the estimated total population from the P Sample; and
M = the estimated number of persons from the P-sample population who match to the census.

Table 9 Estimates of erroneous enumerations missed by the A.C.E. according to the MER and person duplication study. A combined effect is estimated for duplicates without any adjustment for the efficiency of computer matching, and with a 75.7% adjustment applied uniformly. Estimates are based on the MER review subsample. Race and ethnicity categories are based on the A.C.E. poststratification. (Standard errors in parenthesis)

	NonHisp Blacks	Hispanics	All Others	Total
MER only	186,028 (58,056)	221,241 (79,329)	1,047,646 (165,010)	1,454,915 (193,124)
Addl dups, unadj	171,635 (56,172)	211,684 (81,982)	742,058 (170,858)	1,125,377 (199,334)
Total, unadj	357,663 (75,842)	432,925 (110,287)	1,789,704 (248,953)	2,580,292 (283,738)
Addl dups, adj	226,674 (74,184)	279,566 (108,271)	980,017 (225,648)	1,486,257 (263,255)
Total, adj	412,702 (88,612)	500,807 (129,969)	2,027,663 (292,376)	2,941,172 (333,454)
As percent of total E-sample population				
MER only	.63 (.19)	.71 (.25)	.52 (.08)	.55 (.07)
Addl dups, unadj	.58 (.19)	.68 (.25)	.37 (.08)	.43 (.08)
Total, unadj	1.20 (.25)	1.39 (.33)	.89 (.12)	.98 (.11)
Addl dups, adj	.76 (.25)	.89 (.33)	.49 (.11)	.57 (.10)
Total, adj	1.39 (.29)	1.60 (.39)	1.01 (.14)	1.12 (.12)

Although the A.C.E. applied the dual system estimator in 416 separate poststrata, collapsed from an original 448, the effect of the additional erroneous enumerations can be approximated by its effect on a single-cell dual system estimate. Nationally, *DD* = 265,580,677 and N_e = 264,578,863 (Feldpausch, 2001b, p. 79); the match rate M/N_p = 91.59%, and the correct enumeration rate CE/N_e = 95.28% (Davis 2001, Attachment E, p. 1). Using these reported rates, the single-cell dual system estimate of the A.C.E. universe is approximately 276.28 million. If the correct enumeration rate were revised to 94.17% on the basis of the estimated 2,941,172 additional erroneous enumerations, the comparable estimate would be 273.06 million, a change of approximately 3.2 million.

Table 10 repeats the analysis for the age/sex classification used by A.C.E. poststratification. The additional effect of duplicates on 18-29 Males or 18-29 Females is less than the effect measured by the MER. The combined effect of MER plus the duplicates is particularly large for 18-29

Females and small for 30-49 Females, although these differences are principally due to the MER contribution rather than the addition from the duplicate study. Nearly 40% of the estimated total from additional duplicates is contributed by the 0-17 age group, suggesting that this group merits specific study. (The standard error of the estimate is large, however.)

Table 10 Estimates of erroneous enumerations missed by the A.C.E. according to the MER and person duplication study. A combined effect is estimated for duplicates without any adjustment for the efficiency of computer matching, and with a 75.7% adjustment applied uniformly. Estimates are based on the MER review subsample. (Standard errors in parenthesis)

	Total	0-17	18-29 Male	18-29 Female
MER only	1,454,915 (193,124)	249,051 (676,16)	221,210 (62,569)	319,996 (79,552)
Addl dups, unadj	1,125,377 (199,334)	432,629 (107,583)	70,321 (37,663)	115,716 (56,039)
Total, unadj	2,580,292 (283,738)	681,680 (122,295)	291,531 (73,183)	435,712 (89,259)
Addl dups, adj	1,486,257 (263,255)	571,362 (142,082)	92,871 (49,741)	152,823 (74,009)
Total, adj	2,941,172 (333,454)	820,413 (152,276)	314,081 (801,16)	472,819 (99,109)
As percent of total E-sample population				
MER only	.55 (.07)	.36 (.10)	1.17 (.33)	1.55 (.38)
Addl dups, unadj	.43 (.08)	.63 (.16)	.37 (.20)	.56 (.27)
Total, unadj	.98 (.11)	1.00 (.18)	1.54 (.39)	2.11 (.43)
Addl dups, adj	.57 (.10)	.84 (.21)	.49 (.26)	.74 (.36)
Total, adj	1.12 (.12)	1.20 (.22)	1.66 (.43)	2.29 (.48)

Table 10: Combining MER and duplication study results (cont.).

	30-49 Male	30-49 Female	50+ Male	50+ Female
MER only	211,016 (70,420)	47,414 (38,068)	192,607 (70,452)	213,622 (74,814)
Addl dups, unadj	170,966 (71,114)	118,977 (47,137)	140,672 (63,735)	76,096 (46,707)
Total, unadj	381,982 (99,857)	166,391 (60,491)	333,279 (107,577)	289,718 (87,480)
Addl dups, adj	225,790 (93,919)	157,130 (62,253)	185,782 (84,173)	100,498 (61,685)
Total, adj	436,806 (117,135)	204,544 (72,862)	378,389 (124,148)	314,120 (96,103)
		As percent of total E-sample population		
MER only	.54 (.18)	.11 (.09)	.58 (.21)	.53 (.18)
Addl dups, unadj	.44 (.18)	.29 (.11)	.42 (.19)	.19 (.12)
Total, unadj	.98 (.25)	.40 (.15)	1.00 (.32)	.72 (.22)
Addl dups, adj	.58 (.24)	.38 (.15)	.56 (.25)	.25 (.15)
Total, adj	1.12 (.30)	.49 (.18)	1.13 (.37)	.78 (.24)

6. DISCUSSION

The combined evidence from the MER review sample and the person duplication study indicates that the A.C.E. substantially underestimated erroneous enumerations. The actual effect is likely to be even greater than that estimated here, because optimistic estimates for the efficiency of computer matching were used, and this preliminary analysis was restricted to duplicates with weights > .98 only. Furthermore, the MER data showed that the A.C.E. substantially understated some components of erroneous enumeration, and the duplicate study suggests that even the MER understates the magnitude of this understatement. The combined estimates here only use directly identified duplicates and estimates of undetected duplicates due to the efficiency of computer matching, but the results suggest that the MER could also underestimate the contribution to erroneous enumerations of persons enumerated only at the wrong residence and not duplicated.

To use this evidence to refine the A.C.E. estimates and, in turn, to adjust the census estimates for postcensal estimation will require additional research. Practically speaking, the reinterview aspect of the MER cannot be redone, but an expansion of the review subsample can be considered for clerical review in order to refine the MER review estimates used in this report.

Although beyond the scope of this report, the quality of MER data for P-sample residence could be investigated simultaneously. Reclassification of matched E-sample cases as erroneous enumerations would have implications for the P-sample estimates, and these have not yet been studied.

The large number of MER unresolved and conflicting cases leaves the findings of the MER review somewhat unclear, and a more complete analysis of the missing data is important. As noted previously, A.C.E. applied imputation rules for those with another census day address when the address was not reported. The rules resulted in a high rate of imputed erroneous enumerations for these cases; a similar approach could be examined for similar unresolved EFU 2/Best Codes.

The relatively high proportion of duplicates found among the conflicting cases in MER suggests that a high proportion of conflicting cases may represent erroneous enumerations. Further empirical evidence and analysis are required.

Because of its potential to validate and supplement the MER measurement, additional research on methods to identify person duplication now appears warranted. Further work on the Poisson model, including incorporation of the effect of middle initial, merits attention. The model can be extended from national application to matches within the same county or state. A more nuanced approach to using the Census Bureau’s person matching algorithm also appears worth pursuing—the current approach measured the score produced by the algorithm against a single threshold, but then rejected many of the matches found at that threshold. The efficiency of computer matching to detect duplicates, represented here by a uniform 75.7%, requires further empirical investigation.

The estimation procedures used in this analysis easily admit of a number of refinements. Only the MER review sample is used in the analysis, ignoring the information from the rest of the A.C.E. sample. Ratio estimation to results from the person duplication study for the full A.C.E. sample could be considered, as well as other means of combining the MER and duplication information.

References

Adams, Tamara and Krejsa, Elizabeth A. (2001), “ESCAP II: Results of the Person Followup and Evaluation Followup Forms Review,” Executive Steering Committee on Accuracy and Coverage Evaluation Policy II, Report Number 24, DSSD Census 2000 Procedures and Operations Memorandum Series #T-17, Oct. 16, 2001, U.S. Census Bureau.

Bean, Susanne (2001), “ESCAP II: Accuracy and Coverage Evaluation Matching Error,” Executive Steering Committee on Accuracy and Coverage Evaluation Policy II, Report Number 7, Oct. 3, 2001, U.S. Census Bureau.

Bryne, Rosemary; Imel, Lynn; Ramos, Magdalena; and Stallone, Phawn (2001), “Accuracy and Coverage Evaluation: Person Interviewing Results,” DSSD Census 2000 Procedures and Operations Memorandum Series B-5, Feb. 28, 2001, U.S. Census Bureau.

Cantwell, Patrick J., and Childers, Danny R. (2001) “Accuracy and Coverage Evaluation Survey: A Change to the Imputation Cells to Address Unresolved Resident and Enumeration Status,” DSSD Census 2000 Procedures and Operations Memorandum Series Q-44, Mar. 6, 2001, U.S. Census Bureau.

Cantwell, Patrick J.; McGrath, David; Nguyen, Nganha; and Zelenak, Mary Frances (2001), “Accuracy and Coverage Evaluation: Missing Data Results,” DSSD Census 2000 Procedures and Operations Memorandum Series B-7, Feb. 28, 2001, U.S. Census Bureau.

Davis, Peter P. (2001), “Accuracy and Coverage Evaluation: Dual System Estimation Results,” DSSD Census 2000 Procedures and Operations Memorandum Series B-9, Feb. 28, 2001, U.S. Census Bureau.

Fay, Robert E. (2001), “The 2000 Housing Unit Duplication Operations and Their Effect on the Accuracy of the Population Count,” presented at the Joint Statistical Meetings, Atlanta, GA, Aug. 5-9, 2001, and to appear in the *Joint Statistical Meetings 2001* on CD-ROM, American Statistical Association, Alexandria, VA. Two tables omitted from the CD version are available from the author (asa01full.pdf).

Feldpausch, Roxanne (2001a), “Census Person Duplication and the Corresponding A.C.E. Enumeration Status,” Executive Steering Committee on Accuracy and Coverage Evaluation Policy II, Report Number 6, DSSD Census 2000 Procedures and Operations Memorandum Series #T-16, Oct. 12, 2001, U.S. Census Bureau.

Feldpausch, Roxanne (2001b), “E-Sample Erroneous Enumeration Analysis,” Executive Steering Committee on Accuracy and Coverage Evaluation Policy II, Report Number 5, DSSD Census 2000 Procedures and Operations Memorandum Series #T-11, Oct. 10, 2001, U.S. Census Bureau.

Hogan, Howard (1993), “The 1990 Post-Enumeration Survey: Operations and Results,” *Journal of the American Statistical Association*, **88**, 1047-1060.

Hogan, Howard (2001), “Accuracy and Coverage Evaluation: Data and Analysis to Inform the ESCAP Report,” DSSD Census 2000 Procedures and Operations Memorandum Series B-1, March 1, 2001, U.S. Census Bureau.

Krejsa, Elizabeth A. (2001), “ESCAP II: A.C.E. Erroneous Enumerations Errors: Analysis of Census Discrepant Persons,” Executive Steering Committee on Accuracy and Coverage Evaluation Policy II, Report Number 4, Sep. 21, 2001, U.S. Census Bureau.

Martin, Betsy (2001a), “Instrument Differences and their Possible Effects, Comparison of the Evaluation Followup (EFU) and the Person Followup (PFU) Instruments,” internal document, Sep. 17, 2001, U.S. Census Bureau.

Martin, Betsy (2001b), “Instrument Differences and their Possible Effects, Comparison of the Evaluation Followup (EFU) and the Person Followup (PFU) Instruments,” internal document, Oct. 17, 2001, U.S. Census Bureau. (This revision differs from the original primarily in an attached table summarizing the findings.)

Mule, Thomas (2001), “Person Duplication in Census 2000,” Executive Steering Committee on Accuracy and Coverage Evaluation Policy II, Report Number 20, DSSD Census 2000 Procedures and Operations Memorandum Series Q-71, Oct. 11, 2001, U.S. Census Bureau.

Nash, Fay (2000), “Overview of the Duplicate Housing Unit Operations,” internal document, November 7, 2000, U.S. Census Bureau.

Raglin, David A. (2001), “ESCAP II: Effect of Excluding Reinstated Census People from the A.C.E. Person Process,” Executive Steering Committee on Accuracy and Coverage Evaluation Policy II, Report Number 13, Oct. 9, 2001, U.S. Census Bureau.

Notes:

(1) Mule (2001, p.9) reports a weighted estimate of 41,046 duplicates missed clerically out of 724,687 computer matches within cluster, or about 5.7%. On the basis of clerical review, Bean (2001, p. 21) obtained an unweighted estimate of missed duplicates of about 3.3%, largely offset by a 2.4% rate of false duplicates.

(2) As noted in section 3, the source file for the study included both E-sample eligible housing units in the A.C.E. universe and reinstated housing units, which were not in the universe. Table 2 (Mule 2000, p. 9) and Table 3 (p. 11) allow the reinstated units to be partially separated out. The estimate 660, 094 = 660,189 (from Table 3) - 95 (Reinstated to group quarters within cluster, from Table 2). Consequently, the estimate may contain a small number of duplicates of reinstated to group quarters outside surrounding blocks. A similar problem affects Table 4.

(3) Mule (2001) showed that the decline in *duplicates* could largely be attributed to the effect of the Housing Unit Duplication Operations, which had the effect of permanently removing some preliminary census enumerations that appeared to be duplicate housing units and of isolating other units, the reinstated cases,

from the A.C.E. universe, although the reinstated cases were included in the final census. Krejsa (2001) found no significant error in the estimation of *fictitious.* The change in *geocoding* was small and also was investigated with the MER. There were some changes in the standards for *insufficient information* between 1990 and 2000 (Feldpausch 2001, p. 16). The identification of cases as *insufficient information* was governed by rules that could have been implemented clerically without expected difficulty. The MER study allowed reclassification of cases by reapplying the definition, and a small number of cases were reclassified in this way, primary reclassifying some cases out of *insufficient information* into some other status (personal communication, Susanne Bean, Oct. 19, 2001). The effect is included in MES estimates (Bean 2001) but not separately identified.

(4) Martin (2001b, p. 6) identifies situations for which EFU did not obtain addresses: 1) person died after census day/stayed at another residence on census day, 2) person never lived at sample address, 3) college student stayed at another housing unit (HU) on April 1, 4) person moved out before April 1 or in after April 1. She also noted that the EFU did not resolve which residence was correct for persons living in group quarters where persons are allowed to indicate their usual home is elsewhere. These group quarters include, for example, military barracks or ships, but not college dormitories or nursing homes.

(5) Martin (2001a, pp. 1-2) states "It is important to note that interviewers in the Person Followup and the Evaluation Followup were instructed to write notes, and these notes (as well as other auxiliary information) were used heavily in classifying match codes. Even if information would seem to be lacking because there is no pertinent question, it may be available in notes obtained by well-trained interviewers. However, as a rule notes would be considered to provide less reliable and uniform data that [than] responses to standardized survey questions."

(6) Besides *group quarters,* Table 3 (Adams and Krejsa 2001, p.9) includes the categories *movers* (primarily those who would have moved in after Census Day), *never lived here, address mixup, birth or death, other residence--interview at first home, other residence--interview at second home,* and *other residence--unspecified,* which all pertain to *other residence* issues. Persons born after census day or dying before are included in the broad category of *other residence*; the estimate 39,395 (Table 3) indicates that this issue is at most a small component of the underestimation of erroneous enumerations in the E sample. Table 3 does not provide more detail on 76,262 classified as *other*; in fact, some or all may be categorized as *other residence* rather than *fictitious* or *geocoding error*.

(7) Hogan (1993, p. 1056) remarks on the group "other counting errors" which are essentially equivalent to *other residence:* "Most of the "other counting errors" are enumerations of people who moved into the address after Census Day. If they were missed at the correct location, then this may be the only place that they were enumerated. This type of error is often, but not always, paired with census omissions of the actual Census Day residents."

(8) Most links were pairs, but the generalization of the rules to matches among more than two also retained a single unit in the A.C.E. universe and provisionally removed the rest. In determining the unit to retain, an occupied unit was selected over a vacant, the unit with more persons was selected, or when the number of persons was equal, the earlier or more completely recorded unit was selected (Nash 2000 and material it cites).

(9) Matching 1) allowed for reversal of first and last name, 2) removed "Jr," "Sr," and "III" from first and last name fields, 3) located middle initials at the end of names such as "SMITHL" 4) allowed a match of year within one year (Mule 2001, p. 4, App. B).

(10) Most but not all of these names were Spanish; Patrick on March 17 is a familiar non-Spanish example (Mule 2001, App. H). Instances of "John Doe" or "Jane Doe" were also eliminated, as were name/birth date combinations used in census training manuals (Mule 2001, pp J-2 to J-3).

(11) For purposes of replication, the unrounded value of .75718900, based on a preliminary version of Table 5, is included in subsequent calculations. The parallel calculation with the current values in Table 5 gives .75719059.

(12) A few MER records matched to more than one duplicate record. A single duplicate record was selected, with preference for E-sample eligible units first and then for group quarters. Within each type, the record with the highest model weight was chosen.

(13) The weights are related to probabilities, and weight > .98 is used to identify almost certain duplications. For common names, however, where coincidental sharing of birthday is more frequent than census duplications, the weights can be negative and are typically closer to 0 than to 1. To the extent that the weights can be used as effectively probabilities for most names, then introduction of the weights into the analysis could be based on a modification to eq. (5), estimating the first term on the right-hand side with the weights, but omitting the weights in calculating the remaining terms that are subtracted.